ASA Methods Manual (version 0.1)

2023-09-12

Table of contents

Preface

This manual represents the methods used to develop the American Society of Anesthesiologists (ASA) practice parameters. It describes processes, procedures, and relevant policies overseen by the Committee on Practice Parameters (CPP).

As the methods and approaches evolve, modification are incorporated. Those representing ASA policy or falling under CPP's authority are included only after administrative approval (eg, matters related to conflict of interest or the choice of strength of evidence framework). Other changes, for example evidence synthesis methods, are the purview of methodologists. They are updated as appropriate or when clarifications are necessary. A history of substantive modifications are listed at the end of each chapter (in the online version only).

Comments, suggestions for additions, or corrections can be sent to Mark Grant.

1 Introduction

1.1 Background

Practice parameters are “strategies for patient management developed by the profession to assist physicians in clinical decision making” (Health Subcommittee Hearing, 1990). The methods described here apply to the development of ASA Practice Guidelines and Practice Advisories. They are similar in approach and methodologies but differ in that the evidence included in Advisories is limited in overall quantity, quality, and consistency. Classifying a guidance document as a Practice Advisory is accordingly based on the supporting systematic review. Differences notwithstanding, both types of guidance adhere to standards for trustworthy clinical practice guidelines (Graham, 2011).

The first ASA Practice Guidelines, published in 1993, included management of the difficult airway (Caplan et al., 1993) and pulmonary artery catheterization (Roizen et al., 1993). Initial guideline development followed an approach outlined in the Manual for Clinical Practice Guideline Development (Woolf, 1991) commissioned by the Agency for Health Care Policy and Research[1]. The approach was state of the art for guideline development at the time detailing 59 steps accompanied by worksheets, table formats, meeting schedules, and goals. While many of those steps became standard practice in ASA guideline development, others were omitted or modified. Over time, some changes to the guideline development process occurred slowly, while others were more frequent, including the strength of evidence ratings (1999, 6 categories;[2] 2009, 5 categories;[3] 2010, 4 categories;[4] 2013, 3 categories[5]).

Following the release of Clinical Practice Guidelines we can Trust (Graham, 2011) from the National Academy of Medicine and Finding What Works in Health Care: Standards for Systematic Review (Eden, 2011), scrutiny of guideline development increased. In that context, the approach and methods outlined here reflect the evolution of the ASA practice parameter enterprise and their adherence to current standards.

[1] Predecessor to the Agency for Healthcare Research and Quality (AHRQ).

[2] Supportive, suggestive, equivocal, insufficient, inconclusive, silent.

[3] A: supportive literature, B: suggestive literature, C: equivocal literature, D: insufficient evidence from literature, Inadequate.

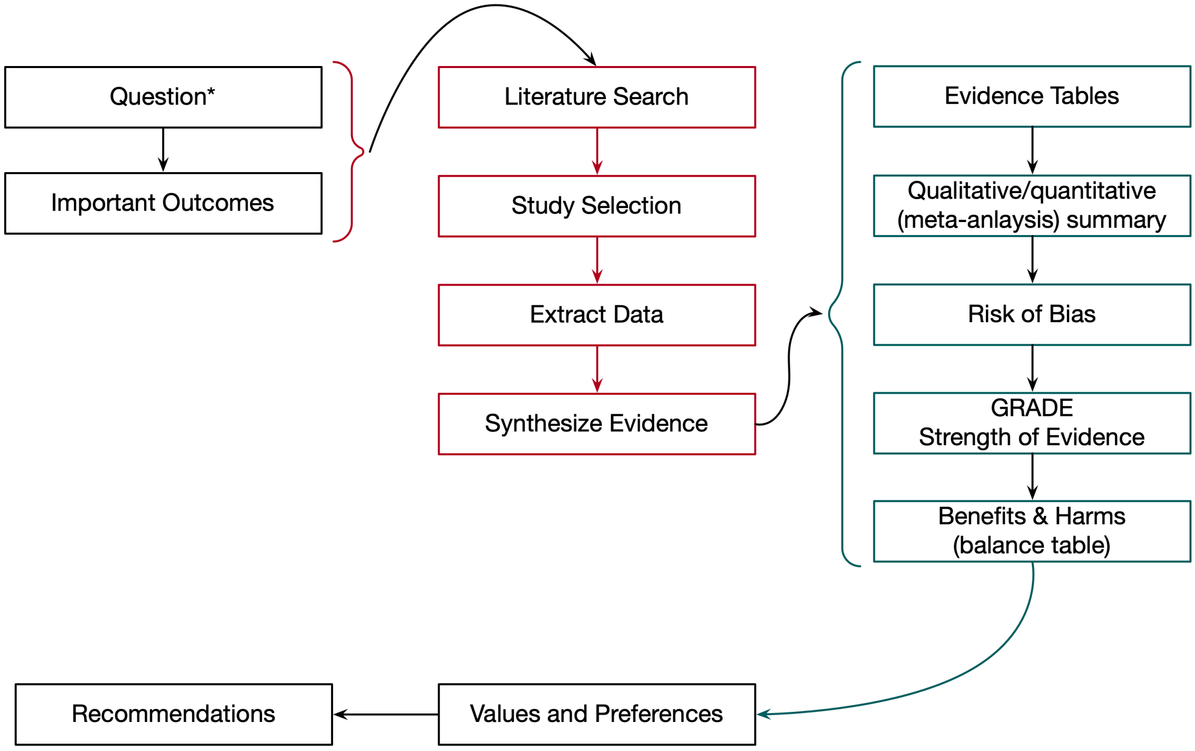
[4] A: supportive literature, B: suggestive literature, C: equivocal literature, D: insufficient evidence from literature.

[5] Category A, Category B, Insufficient Evidence.

1.2 Overview

Figure **??** broadly outlines the structure and main steps followed in developing recommendation for each question and detailed in chapters 3 through # of this manual.

Figure 1.1: The ASA process of developing recommendations.



*Implicit or explicit in each question are a target population, interventions, comparators, outcomes, timing, and setting.

2 Organization

2.1 Committee on Practice Parameters

The Committee on Practice Parameters (CPP) oversees the development of practice parameters, including topic prioritization, reviewing and approving drafts, developing relevant policies (eg, conflict of interest), and evaluating guidelines from other organizations for endorsement[1] or affirmation of value[2]. CPP members are self-appointed and include six active ASA members representing geographically diverse areas, adjunct member(s), and ex officio members from four quality-focused ASA committees. The chair, self-appointed with the ASA president's approval, is responsible for directing and coordinating all committee activities.

2.2 Task Forces

Following a decision to develop a new practice parameter or revise an existing one, the CPP chair forms a task force. A chair (and optional co-chair) leads the task force that includes clinicians, patient representative(s), a librarian/information specialist, methodologists, and the CPP chair. The clinician members are selected based on subject-matter expertise, guideline development and review methodology experience, potential conflicts of interest, and practice diversity. To minimize potential bias across the task force, membership selection seeks diversity in sex, gender, race, ethnicity, practice environment, area of expertise, and geographic region. The task force chair, co-chairs, and CPP chair oversee practice parameter scope, adherence to timelines and ASA methodology.

2.3 Conflict of Interest

Task force members are required to disclose all personal and immediate household member[3] relationships with industry and other entities that might pose a potential conflict of interest.

[1] The document generally satisfies ASA's guideline development requirements, and there is general agreement with all recommendations in the document.

[2] Guideline or practice parameter has merit and value but does not generally satisfy ASA's guideline development requirements, or there is no general agreement with all recommendations in the document.

[3] Partner with whom participant has lived for 1 year in the same home. Dependent or any other related person (by blood or marriage) with whom participant has lived for 1 year in the same home.

Disclosures cover the 3 years preceding the first task force meeting and are updated annually through the year following practice parameter publication. Task force members are asked to avoid as much as possible changes in potential conflicts of interest from the time of appointment to the publication. They must verbally disclose any relevant relationships at the beginning of all conference calls and meetings. Employees of industry, part- or full-time, are prohibited from task force membership.

A task force member has a relevant relationship which is considered a conflict of interest when:

1. The relationship or interest relates to the same or similar subject matter, intellectual property, asset, topic, or issue addressed in the practice parameter.
2. The company/entity with whom the relationship exists makes a drug, drug class, or device addressed by the task force makes a drug or device that competes for use with a product addressed in the practice parameter.
3. The person or household member has a reasonable possibility of financial, professional, or other personal gains as a result of the issues or content addressed by the task force — and is judged to create a risk that a relationship will unduly influence a person's judgment.

Chairs and co-chairs, and at least half of the entire task force (chair, co-chair, other members) must be free of potential conflicts of interest. Task force members without conflicts of interest participate in discussions, drafting, and voting on recommendations. Members with potential conflicts participate in discussions and drafting of documents, but are recused from voting on recommendations related to those conflicts.

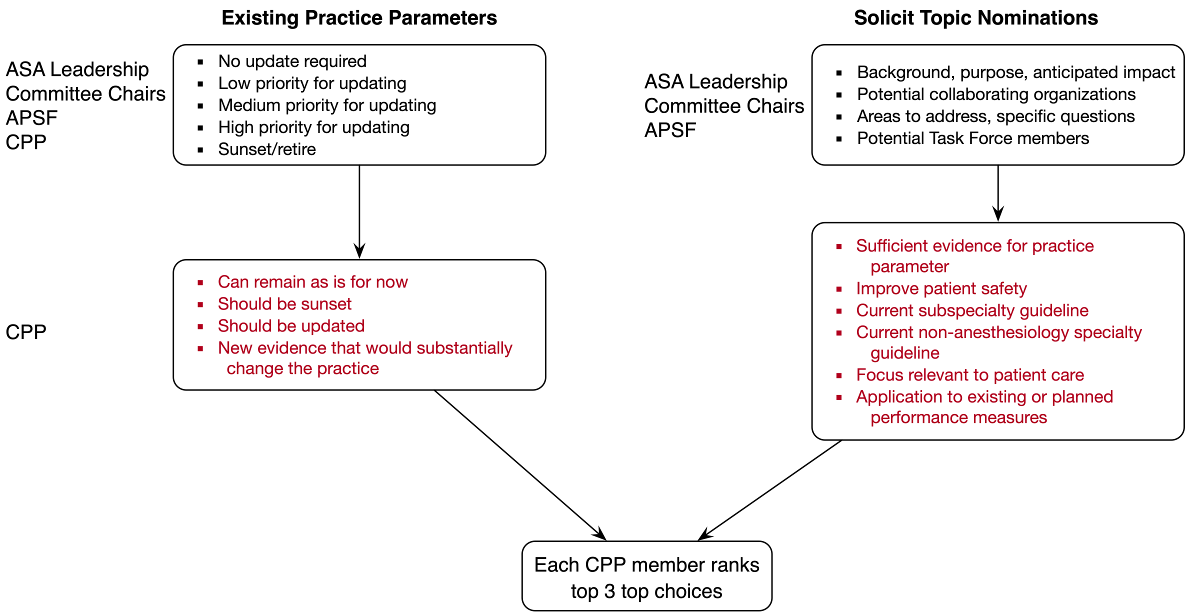
The disclosure policy can be viewed here.

2.4 Practice Parameter Nomination and Prioritization

The process of deciding practice parameters to update or develop is outlined in Figure **??**. Existing practice parameters are prioritized annually for updating by ASA leadership, the Anesthesia Patient Safety Foundation (APSF), committee chairs, and CPP members. Topic nominations are solicited from ASA leadership, committee chairs, and APSF in a standardized format. Nominations for new practice parameters are also accepted from others at any time (sent to the CPP chair or submitted to ASA Standards and Guidelines; see template for suggested content).

Applying evaluation criteria (separate criteria for updating practice parameters and new topics) developed by CPP members, the committee next reviews potential practice parameter updates given the prioritization survey results and new topic nominations. In a final survey conducted following the meeting, each CPP member ranks four top choices.

Figure 2.1: Depiction of the prioritization process.

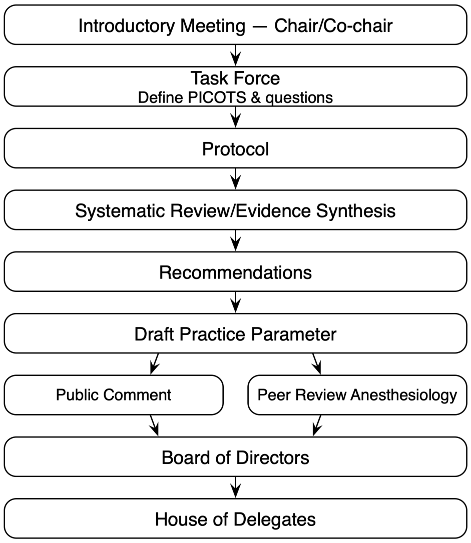

Existing Practice Parameters
Solicit Topic Nominations
ASA Leadership
Committee Chairs
APSF
CPP
No update required
Low priority for updating
Medium priority for updating
High priority for updating
Sunset/retire
ASA Leadership
Committee Chairs
APSF
Background, purpose, anticipated impact
Potential collaborating organizations
Areas to address, specific questions
Potential Task Force members
CPP
Can remain as is for now
Should be sunset
Should be updated
New evidence that would substantially change the practice
Sufficient evidence for practice parameter
Improve patient safety
Current subspecialty guideline
Current non-anesthesiology specialty guideline
Focus relevant to patient care
Application to existing or planned performance measures
Each CPP member ranks top 3 top choices

CPP: Committee on Practice Parameters; APSF: Anesthesia Patient Safety Foundation

2.5 Developing Practice Parameters

Figure **??** outlines the practice parameter development process. An introductory meeting serves to orient the task force chairs and co-chairs to the process, timeline, and the roles of methodologists. Subsequent task force meetings are then devoted to defining the PICOs (populations, interventions, comparators, and outcomes) and key questions questions. A protocol is then drafted by the methodologists and reviewed by the task force 2 to 4 weeks later. The systematic review and evidence synthesis is then conducted, during which time the task force is convened as needed for input and decisions concerning any issues that arise including modifications to the protocol. The methodologists complete the evidence synthesis to inform recommendations. Finally, the practice parameter is drafted, submitted to Anesthesiology for review, public comment is solicited, followed by submission to the ASA Board of Directors for approval and finally the House of Delegates.

Figure 2.2: Depiction of the practice parameter development process.


Introductory Meeting — Chair/Co-chair
Task Force
Define PICOTS & questions
Protocol
Systematic Review/Evidence Synthesis
Recommendations
Draft Practice Parameter
Public Comment
Peer Review Anesthesiology
Board of Directors
House of Delegates

3 Systematic Review

Trustworthy clinical practice guidelines (Graham, 2011) are supported by systematic reviews meeting explicit standards (Eden, 2011; PCORI, 2019). The systematic reviews supporting ASA practice parameters conform to those standards.

3.1 Protocol

The protocol, developed collaboratively between the task force and methodologists, guides systematic review conduct, and provides documentation for updates. It includes background material, key questions, PICOTS,[1] analytic framework, study inclusion and exclusion criteria, search strategy, and the anticipated approach to evidence synthesis. Depending on the anticipated scope, protocols may be registered on PROSPERO (Booth et al., 2012). However, when the systematic review includes numerous questions and anticipated to require substantial refinement and modifications, registration is omitted. The protocol is included as a supplement to the published practice parameter. (An example draft protocol can be viewed here).

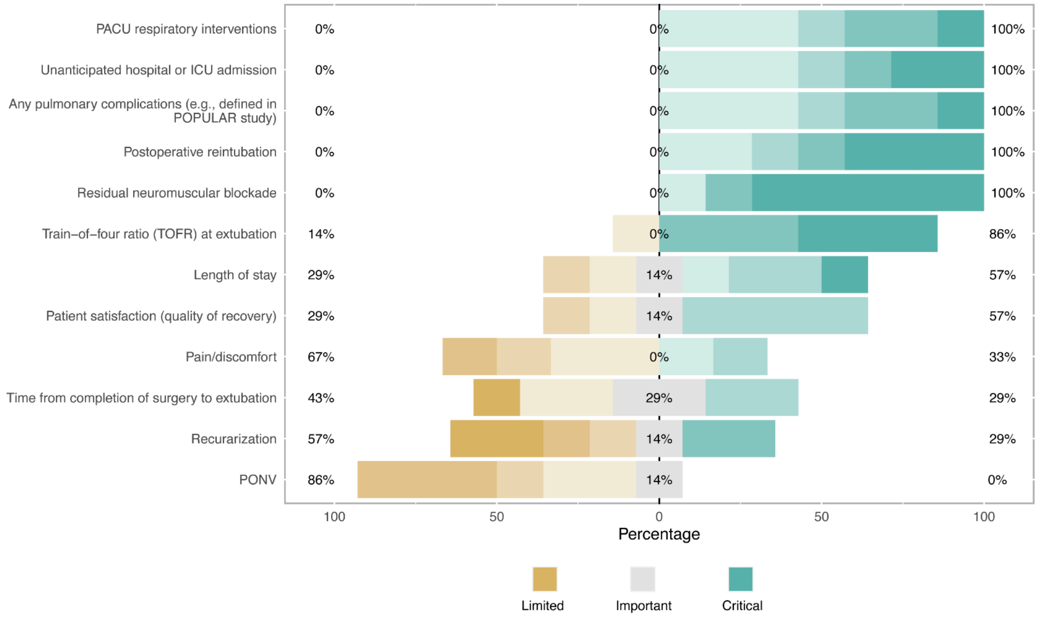
3.2 Outcome Importance

Outcomes vary in importance according to patient values and preferences (Guyatt et al., 2011). From that perspective, following protocol completion task force members rate outcome importance for decision-making. The ratings are reviewed by the entire task force and revised as necessary to achieve consensus. Outcomes are assigned a level — critical, important but not critical, low importance — and may be ranked to prioritize conduct of the evidence synthesis. Figure **??** and Figure **??** illustrate the data obtained from ratings and rankings.

Note that rows can be reordered according to ranking by clicking on column headers.

[1] Populations, interventions, comparators, outcomes, timing, and setting.

Figure 3.1: Prioritization of outcomes for neuromuscular monitoring.



3.3 Identifying Literature

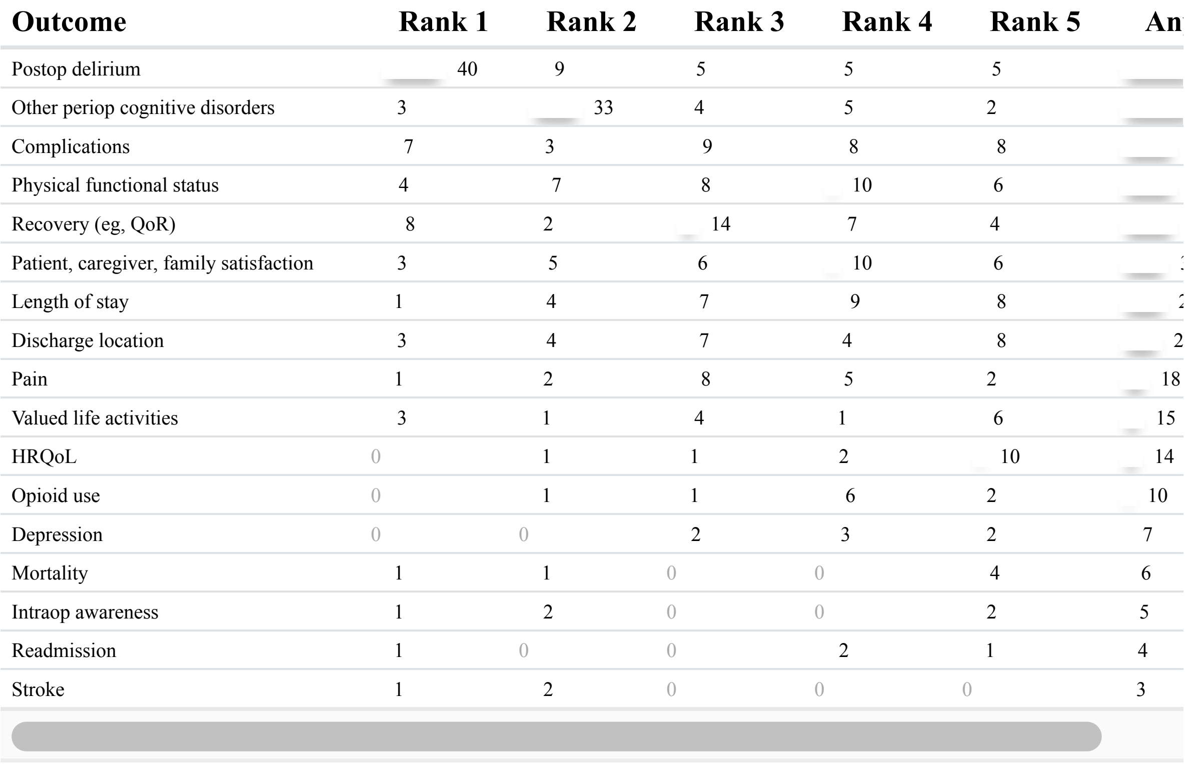
3.3.1 Database Searches

A librarian/information specialist develops search strategies after reviewing the protocol and participating in task force meetings. The primary bibliographic databases queried include PubMed, Embase®, Scopus®, and Cochrane Central Register of Controlled Trials. The task force also submits relevant references for consideration, including systematic reviews and guidelines for reference checking. To ensure that relevant publications have been captured, search result identification of references submitted by the task force is examined. Grey literature searches are topic-dependent relying on registries, conference abstracts, preprint servers, and FDA documents including advisory meeting transcripts. The search dates are determinied by the task force and consider sensitivity (Xu et al., 2022), applicability and generalizability to current practice, and resources required to conduct the review. Depending on the key question, searches may not be limited to English language publications (Egger et al., 1997; Jia et al., 2020; Jüni et al., 2002; Mao et al., 2020).

3.3.2 Citation Searching

Backwards searching for studies included in relevant systematic reviews, meta-analyses, and guidelines are considered eligible for inclusion. The selection process outlined below (Figure **??**)

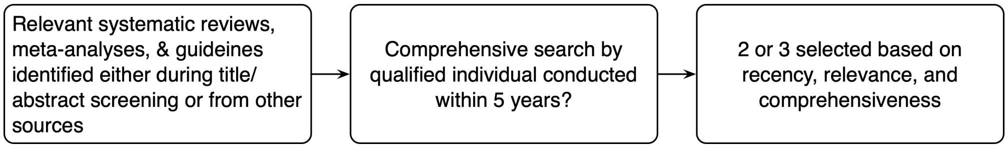
Figure 3.2: Example of assessing outcome importance rankings in a geriatrics guideline. Rankings for the 5 most important outcomes across 7 key questions (11 respondents with maximum 77 for each outcome rank or any top 5 ranking).



Outcome	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	An
Postop delirium	40	9	5	5	5	
Other periop cognitive disorders	3	33	4	5	2	
Complications	7	3	9	8	8	
Physical functional status	4	7	8	10	6	
Recovery (eg, QoR)	8	2	14	7	4	
Patient, caregiver, family satisfaction	3	5	6	10	6	3
Length of stay	1	4	7	9	8	2
Discharge location	3	4	7	4	8	2
Pain	1	2	8	5	2	18
Valued life activities	3	1	4	1	6	15
HRQoL	0	1	1	2	10	14
Opioid use	0	1	1	6	2	10
Depression	0	0	2	3	2	7
Mortality	1	1	0	0	4	6
Intraop awareness	1	2	0	0	2	5
Readmission	1	0	0	2	1	4
Stroke	1	2	0	0	0	3

is used to identify typically 2 to 3 reviews. Studies included in the those reviews are compiled in a bibliographic database. Those studies not identified in the primary search are subsequently assessed for eligibility. On a selective basis, forward citation searching is conducted using seminal studies to identify citing studies. Citationchaser (Haddaway et al., 2022) and/or Paperfecter (Pallath et al., 2023) are used to facility citation searching.

Figure 3.3: Approach to backward citation searching.


Relevant systematic reviews, meta-analyses, & guideines identified either during title/ abstract screening or from other sources
Comprehensive search by qualified individual conducted within 5 years?
2 or 3 selected based on recency, relevance, and comprehensiveness

3.3.3 Task Force

The task force is given the opportunity to submit potentially relevant primary studies, guidelines, systematic reviews, and meta-analyses. The non-primary research are included in the reference checking process and the remainder considered in the standard selection process.

3.3.4 Retracted Publications

Identifying retracted publications is critical to assuring the integrity of the systematic review. Accordingly, searches for retractions of included studies are conducted using relevant search terms (eg, see this guide) and the Retraction Watch Database (can be facilitated using Zotero's Retracted Items feature).

3.3.5 Deduplication

Deduplication is performed using EndNote™ (used as the primary bibliographic database) and a dedicated systematic review software platform (DistillerSR).

3.4 Study Selection

Based on the inclusion-exclusion criteria (study design and PICOTS), study selection is performed by reviewing titles and abstracts. The semi-automated predictive tool for title and abstract screening implemented in DistillerSR is utilized. (Polanin et al., 2019) If the number of references is exceedingly large (eg, > 10,000 or 15,000), screening may be truncated when inclusion predictions for the remaining unscreened references are low (eg, less than 2% to

3%). Full-text review of potentially relevant publications is then conducted with reasons for exclusion at the full-text stage are recorded using a standard set of justifications.

Study designs considered eligible for specific key questions are determined by the questions, PICOTS, and evidence availability. For example, although randomized designs generally offer the most convincing evidence, if few address a particular question/PICOTS, nonrandomized designs may be included. Similarly, nonrandomized designs may be included for evaluation of harms. Case reports and case series, conference abstracts, letters not considered brief research reports, non-English publications, and animal studies are generally not considered eligible.

Two reviewers independently apply inclusion-exclusion criteria at each stage with discrepancies resolved by consensus including a third reviewer. Training sets are used to develop agreement concerning the application of inclusion-exclusion criteria.

3.5 Data Extraction/Management

Accurate data extraction, quality control, and data management enhance reproducibility and support valid evidence synthesis. The workflow standardizes data extraction into a dedicated database with an audit log, and once entered, minimizes manual data manipulation (eg, cutting and pasting).

A standard review-specific set of data entry forms, modified for each systematic review, are used:

- Study characteristics
- Study arm data
- Dichotomous outcomes (as reported)
- Continuous outcomes (as reported)
- Likert or other rating scale outcomes (as reported)
- Risk of bias

Data are abstracted by a single reviewer with verification (PCORI, 2019, pp. SR–1) of data relevant for quantitative synthesis and rating (GRADEing) the strength of evidence. Figures are digitized as necessary to obtain results for synthesis. Data are maintained and edited in DistillerSR, a data dictionary compiled, and then transferred to a local repository for evidence synthesis or reports created using DistillerSR. A sample study characteristics form can be seen here.

3.6 Study Risk of Bias Assessment

Risk of bias for individual studies are evaluated using tools relevant for the study design. The most commonly used tools include:

- Randomized clinical trials — Cochrane risk of bias tool 2.0
- Nonrandomized studies of interventions — ROBINS-I (Risk Of Bias In Non-randomized Studies of Interventions)
- Diagnostic studies – QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies)

Risk of bias assessments are performed independently by 2 reviews and discordances in domain and signalling question results reconciled by consensus including a third reviewer as necessary. Separate risk of bias assessments are conducted for clinical and patient reported outcomes, but may not be conducted for each of the typically multiple outcomes examined.

4 Evidence Synthesis

4.1 Introduction

A single study is rarely sufficient to inform a guideline or policy recommendation[1] (Spiegelhalter et al., 2004, p. 267); a synthesis of evidence obtained from multiple studies is required. The evidence synthesis may be qualitative or quantitative ranging from narrative descriptions of study results to pairwise meta-analysis (a single intervention and comparator) or network meta-analysis (multiple interventions or comparators). Regardless of the approach, the purpose of an evidence synthesis is to summarize benefits, harms, and uncertainty (statistical and non-statistical) to inform decisions and recommendations.

Figure **??** depicts how the evidence synthesis is structured for each key question and how results support recommendations. The figure implicitly emphasizes how guideline users have varied needs with respect to detail. Some are interested only in recommendations that include no quantitative information. Many (hopefully most) seek to understand the the summaries detail in the balance tables. Accordingly, these elements are included in the body of the published guideline. Others may want to understand details including GRADE domains, meta-analyses, and how specifics of the synthesis supports recommendations — provided as supplementary materials (eg, see example). Finally, a rare individual may wish to explore analysis or reproduce them — data and code are made available for that purpose.

It should be noted that explanatory text in the guideline is by necessity more limited than a singular publication devoted to each key question. However, the degree of detail provided should be sufficient to allow a transparent view to the most discerning or critical reader.

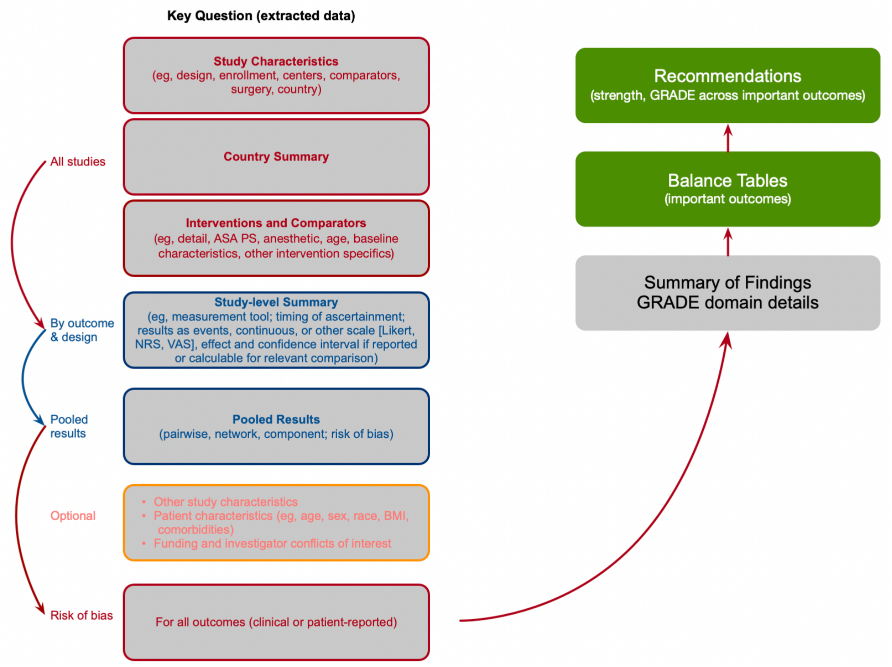
Boxes in gray are included as supplements to the guideline; those in green are part of the publication.

4.2 Decision-Making Frameworks

The decision-making required to develop recommendations requires a framework or model — a calculus of benefits and harms, how they are valued, and their respective uncertainties. The explicitness of the decision calculus varies (Meltzer et al., 2011). For example, a model can be conceptual existing only in the mind of a decision maker with little or nothing quantitative.

[1] "It is unusual for a policy question to be informed by a single study."

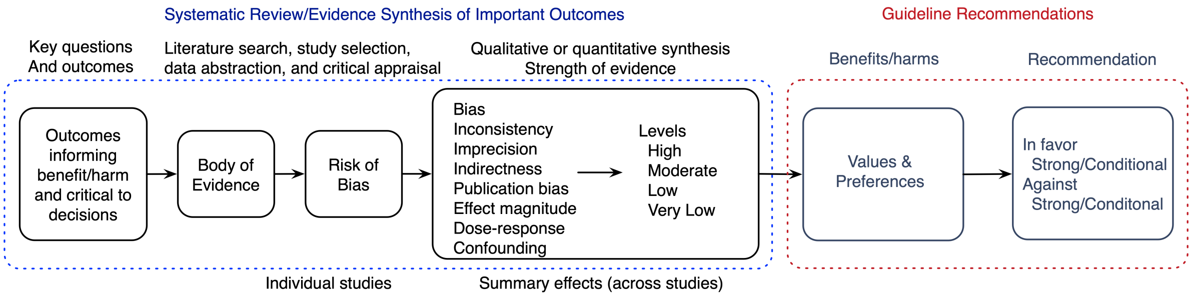
Figure 4.1: Schematic of the evidence synthesis in relation to developing recommendations.


Key Question (extracted data)
Study Characteristics
(eg, design, enrollment, centers, comparators, surgery, country)
Country Summary
Interventions and Comparators
(eg, detail, ASA PS, anesthetic, age, baseline characteristics, other intervention specifics)
Study-level Summary
(eg, measurement tool; timing of ascertainment; results as events, continuous, or other scale [Likert, NRS, VAS], effect and confidence interval if reported or calculable for relevant comparison)
Pooled Results
(pairwise, network, component; risk of bias)
• Other study characteristics
• Patient characteristics (eg, age, sex, race, BMI, comorbidities)
• Funding and investigator conflicts of interest
For all outcomes (clinical or patient-reported)
All studies
By outcome & design
Pooled results
Optional
Risk of bias
Recommendations
(strength, GRADE across important outcomes)
Balance Tables
(important outcomes)
Summary of Findings
GRADE domain details

On the other extreme, the model can decision-analytic with explicit quantitative inputs and outputs. Like most guideline enterprises, the ASA adopts an approach with qualitative and quantitative elements between the extremes.

As outlined in Figure **??**, after formulating key questions and important outcomes specified, relevant studies are identified, data extracted, and risk of bias appraised. Next, using a quantitative or qualitative synthesis, the strength (certainty/GRADE) of evidence for each outcome is rated. Outcomes are then weighted according to patient values and preferences, considered as a whole, and recommendations formulated.

Figure 4.2: Model and approach to evidence synthesis for making recommendations.



But "all models are wrong" (Box, 1976) — including GRADE.

4.3 Quantitative Synthesis

Based on the key question, included studies, clinical and methodological diversity, results are pooled in either pairwise or network meta-analyses. Random effects models are fitted given the goal of estimating unconditional effects (ie, effects not relevant only to the pooled studies) (Hedges et al., 1998). For binomial outcomes, default models use the Mantel-Haenszel method; for continuous outcomes inverse variance weighting. The restricted maximum likelihood estimator is used to estimate between-study variance (Viechtbauer, 2005). For continuous or scale-reported outcomes, if means and standard deviations are unavailable for they are imputed if authors reported medians, interquartile and/or overall ranges for the effects of interest; and if necessary P-values are used to estimate missing standard deviations (Shi et al., 2020). When five or more studies are pooled, the Hartung-Knapp adjustment is applied (Cornell et al., 2014). Network meta-analyses are conducted using frequentist (Balduzzi et al., 2023) or Bayesian methods (Béliveau et al., 2019; Dias et al., 2018) with non-informative priors. Consistency is examined by comparing direct to indirect evidence in the frequentist network meta-analyses and inconsistency models in the Bayesian approach.

Relative effects as reported as risk ratios for clinical interpretability and continuous outcomes as mean differences or standardized mean differences for outcomes reported with differing scales. When feasible, standardized mean differences are re-expressed on the most common scale used. Statistical heterogeneity is examined using the between study variance and I ^{2} (Rücker et al., 2008) and when relevant and practicable explored in subgroup analysis or meta-regression (Schwarzer et al., 2015; Simon G. Thompson et al., 2002) Small-study effects and the potential for publication bias are examined using funnel plots (comparison-adjusted for network meta-analyses), regression-based tests, and adjustment methods (Balduzzi et al., 2019; Harrer et al., 2021). Owing to its statistical properties, results using odds ratios are used reported for when examining small-study effects for relative effects; sensitivity analyses are performed using risk ratios.

Analyses are conducted using R (R Core Team, 2023) in a reproducible manner and made publicly available when the practice parameter is completed.

4.3.1 Harms/Adverse Events

Comparative harms are pooled as either relative or absolute effects according to the frequency of events. For rare events (eg, mortality) risk differences are most often used.

4.3.2 Selected Analysis Matters

Age

Surgical classification

The recovery phases described by these tools can be categorised as early, intermediate and late. The early postoperative recovery phase has been defined as the first 24 h [5, 6] or the first seven days [7–9]. The speed and extent of recovery in the early phase is influenced most by pain, nausea, peri-operative medications and delirium [10]. The intermediate phase of postoperative recovery has been defined as the first 28 [11] or 60 [12] days. The extent of recov- ery in the intermediate phase is influenced most by pain, anxiety and depression, physical impairment and cognitive dysfunction. The late postoperative recovery phase has been defined as the first six weeks [13] or three months [14].

Bowyer 2016

4.3.3 Sensitivity Analyses

Although useful, meta-analytic results are not without limitations (Ioannidis, 2016; Maclure et al., 2001) The robustness of meta-analytic results requires consideration. Aspects include model decisions, small-study effects, influential studies and other factors (eg, secular trends, subgroups).

Modeling decisions include the choice of effect measure, estimators, use of Hartung-Knapp adjustment, continuity corrections for rare events, and even considering studies without events (can be important for harms). As relevant, particularly when uncertainty in a pooled effect appears unclear, how each of these choices impact the range of plausible effects may be examined.

Small-study effects are particularly important as they may represent publication bias or selective reporting and can affect the strength of evidence. *There is, however, no test for publication bias or selective reporting.* The presence of small-study effects offer clues to their potential presence, but require a sufficient number of studies (eg, 10 or more) to effectively examine. Additionally, there are multiple tests (Begg et al., 1994; Egger et al., 1997; Macaskill et al., 2001; Peters et al., 2006; Pustejovsky et al., 2019; Sterne et al., 2011; S. G. Thompson et al., 1999) and adjustment methods — trim and fill (Duval et al., 2000), PET-PEESE (Stanley et al., 2014), limit meta-analyis (Rücker et al., 2011), P-curves (Simonsohn et al., 2014), and selection models (Copas et al., 2014) that can utilized — sometimes offering conflicting results.

We adopt a pragmatic approach to sensitivity analyses. If a pooled result is obtained from 10 or more studies, the most appropriate choice of a regression-based test is reported. But if more than one could be used and results differ both are noted. For adjustment, a limit meta-analysis is our method of choice superimposed on a funnel plot (others may be used as sensitivity checks). Although the Hartung-Knapp adjustment is applied given 5 or more studies, we recognize that published meta-analyses typically do not. It is therefore important to understand sensitivity of results to its use and note in the interpretation of evidence and GRADEing.

Finally, despite the wide range of analytic choices our perspective is that a convincing body of evidence should not be materially impacted for a strong recommendation. If it is, then consideration must be incorporated in both rating the strength of evidence and recommendation.

4.4 Rating the Strength of Evidence

The strength (certainty) of evidence for important outcomes is appraised using the Grades of Recommendation, Assessment, Development, and Evaluation (GRADE Schünemann et al., 2013) and American College of Cardiology/American Heart Association (ACC/AHA Halperin et al., 2016) frameworks.

Different conceptual models (Spiegelhalter et al., 2011) underpin these strength of evidence frameworks — certainty of evidence (GRADE) and evidence hierarchy (ACC/AHA). The longstanding evidence hierarchy (pyramid) model asserts that systematic reviews and meta-analyses of randomized clinical trials provide the most convincing evidence followed by randomized clinical trials, observational studies, and case series or case reports. The certainty of evidence model incorporates the hierarchy insofar as it reflects study validity, but defines strength of evidence in terms of how convinced reviewers are that estimates are close to some "true effect".

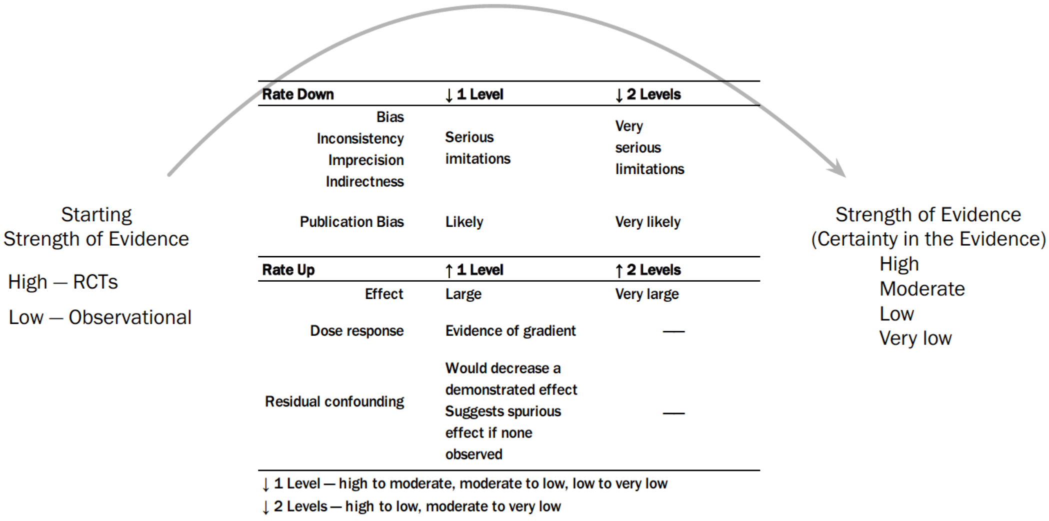
4.4.1 GRADE

Table 4.1: GRADE levels of evidence.

GRADE	Definition
High	We are very confident that the true effect lies close to that of the estimate of the effect.
Moderate	We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different.
Low	Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect.
Very low	We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect.

In the GRADE approach, a strength of evidence is determined using an algorithm that includes limitations in the body of evidence (bias, inconsistency, imprecision, indirectness, publication bias) and factors that can increase confidence in effects[2] (large or very large effect magnitude, dose-response, and plausible residual confounding). According to study limitations, the strength of evidence may be rated down 1 or 2 levels according to study limitations from a starting rating of high for RCTs. Evidence from observational studies begin with a low rating and may be rated down for limitations or rated up because of effect magnitude, dose-response, or the impact of plausible residual confounding. GRADE guidance for rating the certainty of evidence up or down is followed, with some additions. Inconsistency (unexplained heterogeneity) of pooled effects is judged by examining statistical measures (I^2 and between-study variance τ^2) alongside prediction intervals when there are sufficient studies. Owing to its well-described limitations, I^2 and some categorization of it's magnitude (eg, small, moderate, or large), is not used as the primary determinant of heterogeneity (Rücker et al., 2008). These statistics can vary by effect (associational) measures (eg, risk vesrus odds ratios and mean versus standardized mean differences) and are examined for differences in choice.

Figure 4.3: Process of GRADEing the strength (certainty) of evidence (after Balshem et al., 2011).



Starting
Strength of Evidence

High — RCTs

Low — Observational

Rate Down	↓ 1 Level	↓ 2 Levels
Bias Inconsistency Imprecision Indirectness	Serious imitations	Very serious limitations
Publication Bias	Likely	Very likely
Rate Up	**↑ 1 Level**	**↑ 2 Levels**
Effect	Large	Very large
Dose response	Evidence of gradient	—
Residual confounding	Would decrease a demonstrated effect Suggests spurious effect if none observed	—

↓ 1 Level — high to moderate, moderate to low, low to very low
↓ 2 Levels — high to low, moderate to very low

Strength of Evidence
(Certainty in the Evidence)
High
Moderate
Low
Very low

4.4.1.1 Note on Nonrandomized Designs, ROBINS-I, and GRADE

In 2019, the GRADE working group offered "guidance regarding how systematic review authors, guideline developers, and health technology assessment practitioners using GRADE might approach the use of ROBINS-I as part of the certainty rating process" (Schünemann et al., 2019). They suggested that owing to differences in the ROBINS-I tool (referent to trial emulation) that when it is used, the GRADE for a body of evidence from non-randomized

[2] Applies primarily to evidence obtained from observational studies.

designs should start at high, not low. Although a rationale is offered (potential for additional down GRADEing due to confounding and selection bias), our view is that this guidance is not appropriate. No matter how careful the conduct and analysis of a non-randomized design the assumption of no unmeasured confounding is unverifiable (Schulz et al., 2023) (in stark contrast to randomized designs). On this basis, it is logically inconsistent to equate randomized and non-randomized designs. Additionally, the guidance introduces dependence of GRADE on the risk of bias tool — absent in its original formulation. We choose to review carefully our appraisals of nonrandomized evidence to avoid overzealous down GRADEing, but retain GRADE in its original formulation.

4.4.1.2 Comment on GRADE

GRADE is complex. The web-based handbook spans over 40,000 words and there are now over 30 explanatory publications. The 4-level quality of evidence ratings imposes cutoffs for what is in reality a continuous scale — ratings near the cutoffs are less certain than might be evident. We sometimes struggle assigning a GRADE. Accordingly, the categorical GRADE (high, moderate, low, very low) reflects in effect the mode of a distribution. Uncertainty is not conveyed see Llewellyn et al. (2015) and Stewart et al. (2015).

4.4.2 ACC/AHA

Table 4.2: ACC/AHA levels of evidence.

Level	Definition
A	High-quality evidence from more than 1 RCTs. Meta-analyses of high-quality RCTs. One or more RCTs corroborated by high-quality registry studies.
B-R	Moderate-quality evidence from 1 or more randomized controlled trials. Meta-analyses of moderate-quality RCTs.
B-NR	Moderate-quality evidence from 1 or more well-designed, well-executed nonrandomized studies, observational studies, or registry studies. Meta-analyses of such studies.
C-LD	Randomized or nonrandomized observational or registry studies with limitations of design or execution. Meta-analyses of such studies. Physiological or mechanistic studies in human subjects.
C-EO	Consensus of expert opinion based on clinical experience when evidence is insufficient, vague, or conflicting.

RCT: randomized clinical trial; NR: nonrandomized; LD: limited data; EO: expert opinion

ACC/AHA ratings are based on an evidence hierarchy approach. Its 5-level scheme (A, B-R, B-NR, C-LD, C-EO) considers study design (RCT, observational, mechanistic) and quality, number of studies, meta-analytic results, expert opinion and clinical experience (Table **??**). The framework explicitly allows for consideration of mechanistic studies conducted in humans (Goodman et al., 2013). Unlike GRADE guidance for arriving at a strength of evidence is limited to the level of evidence definitions. The ratings were recently revised from an A, B, C scheme by expanding B and C into 2 subcategories and adding E for expert opinion (A. K. Jacobs et al., 2014).